Repetitions in Partial Words

Robert Mercaș

Christian-Albrechts-Universität zu Kiel

Bucharest, August 2013

Table

01010100 01100001 01100010 01101100 01100101

Partial motivation

Noise

DNA alignment

Approximate matching

Partial words

•arte

carte

parte

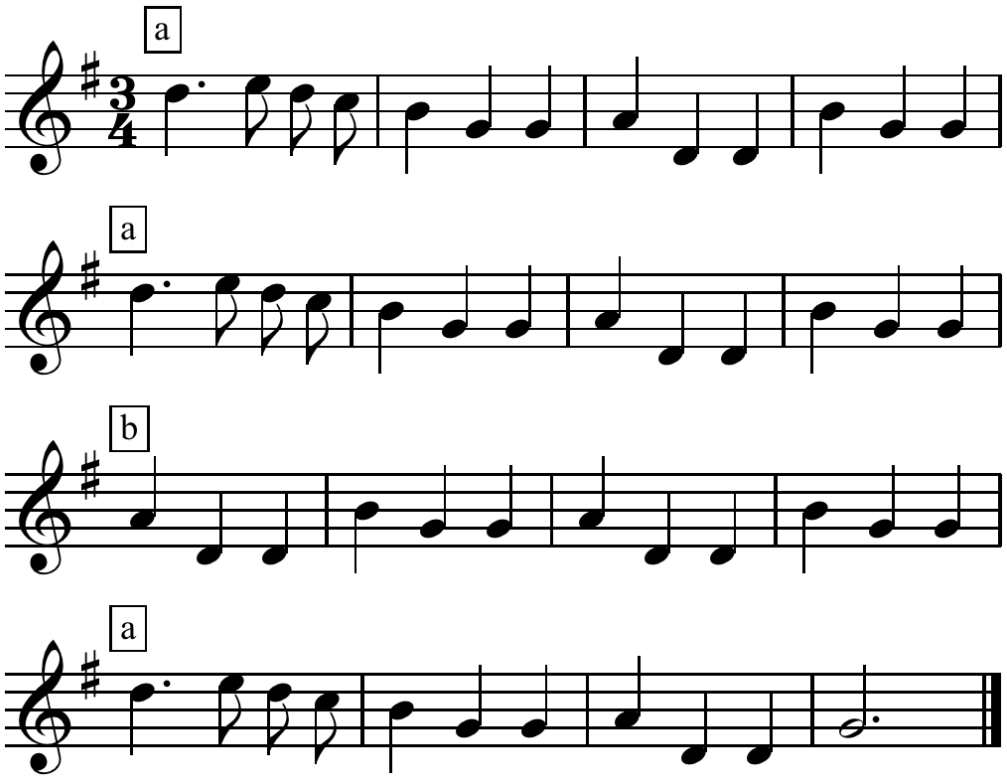
Marte

tarte

sarte

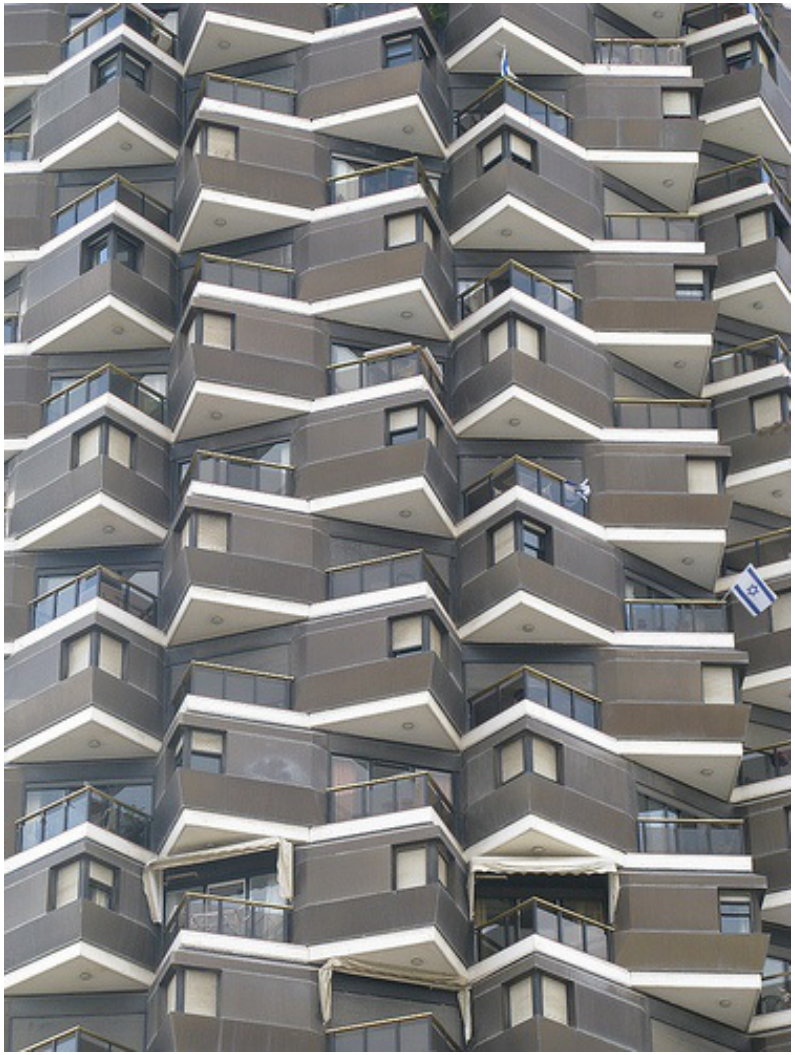
Music

"Repetition, when done skilfully by a master composer, is emotionally satisfying to our brains, and makes the listening experiences as pleasurable as it is." [Levitin 07]



Oh du lieber Augustin

Other repetitions



Word repetitions

Roentgenium (element with atomic number 111) was called

ununun**ium**

3-repetition (cube) of the factor *un*

anana**s**

$\frac{5}{2}$-repetition of the factor *an*

Motivation

- **Combinatorics on words**
  Avoidability of repetitions, Interaction between periods, Counting repetitions
- **Pattern matching algorithms**
  String Matching, Time-space optimal String Matching: local and global periods, Indexing
- **Text Compression**
  Generalised run-length encoding, Dictionary-based compression
- **Analysis of biological molecular sequences**
  Intensive study of satellites, Simple Sequence Repeats, or Tandem Repeats in DNA sequences, Molecular structure prediction
- **Analysis of music**
  Rhythm detection, Chorus location

Full Words

Alphabet: A – a non-empty finite set
Letter/Symbol: $a \in A$
Word: $w = a_0 \ldots a_{n-1}$ – a finite concatenation of symbols $a_i \in A$
Length of w: $|w|$ – the number of symbols in w
Empty word: ε – unique word of length zero

The word *abaababbbba* is defined over the alphabet $A = \{a, b\}$ and has length 11.

Partial Words

A *partial word* is a sequence of symbols over a finite alphabet that may contain a number of "do not know" symbols or "holes".
A hole will be denoted by $\diamond$.
Alphabet: $A_\diamond$ – where $\diamond \notin A$

If $|w|_\diamond = 0$ then w is a full word.

The partial word $aba\diamond bab\diamond\diamond ba$ is defined over the alphabet $A = \{a, b\}$ and has length 11.

For $u, v \in A_\diamond^*$:

If $|u|=|v|$ then

- u is *contained* in v ($u \subset v$), if $u(i)=v(i)$ for all i with $u(i)\in A$.

$$\diamond abba\diamond \not\supset a\diamond bba\diamond \subset aabbab$$

- u, v are *compatible* ($u \uparrow v$), if $w\in A_\diamond^*$ exists with $u\subset w$ and $v\subset w$.

$$a\diamond bba\diamond \subset aabbab \supset \diamond abba\diamond$$

$$a\diamond bba\diamond \uparrow \diamond abba\diamond$$

If $v = xuy$ then

- u is a *factor* of v.
- u is *proper factor* of v if $u \neq \varepsilon$ and $u \neq v$.
- u is a *prefix* of v if $x = \varepsilon$ and a *suffix* of v if $y = \varepsilon$.

Basic facts

Let $\phi(a) = ab$ and $\phi(b) = ba$ and define $\tau_0 = a$ and $\tau_i = \phi^i(a)$. Remark that $\tau_{i+1} = \phi(\tau_i)$ and $\tau_{i+1} = \tau_i\overline{\tau}_i$, where $\overline{x}$ is the complement of x.

Thue-Morse word

$\tau = \lim_{i\to\infty} \tau_i = \lim_{i\to\infty} \phi^i(a)$ is overlap-free.

abbabaabbaababbabaababbaabbabaababbabaabbaababbabaababbaabbab

Let $\psi(a) = abc$, $\psi(b) = ac$ and $\psi(c) = b$, and define $\sigma_0 = a$ and $\sigma_i = \psi^i(a)$. Remark that $\sigma_{i+1} = \psi(\sigma_i)$.

Hall word

$\sigma = \lim_{i\to\infty} \sigma_i = \lim_{i\to\infty} \psi^i(a)$ is square-free.

abcacbabcbacabcacbacabcbabcacbabcbacabcbabcacbac

Observations

For partial words, whenever considering powers we will refer to the compatibility concept:

$ba\diamond bab\diamond ab\ ba\diamond\ bab\ \diamond ab$ is a cube.

- Every word containing a $\diamond$ has a (trivial) square occurrence ($a\diamond b$).
- If two holes are too close to each other we can obtain high powers ($\diamond a\diamond$ is a cube)

Freeness for partial words

There exist infinitely many partial words with infinitely many holes over a

- ternary alphabet, avoiding all squares; 13 in σ
  $\tau_4(a) = abcacbabcbac\diamond bcacbacabcb$
- ternary alphabet, avoiding all overlaps; 13 in σ
  (only binary alphabet necessary for one hole) $\diamond\tau$
- binary alphabet, avoiding all cubes. 14 in τ
  $\sigma_5(a) = abbabaabbaaba\diamond babaababbaabbabaab$

Idea: replace a position of τ or σ by $\diamond$. (Iterate)

[Blanchet-Sadri, M., Scott, TCS 09] [Manea, M., TCS 07]

Freeness for arbitrary insertion

We want to construct words which preserve some freeness properties even after an arbitrary insertion of holes.

Between two inserted holes we require at least two letters of the alphabet.

There exist infinitely many words which after arbitrary insertion of holes

- avoid all squares, except $a\diamond$, $\diamond a$ and $\diamond ab\diamond$, over eight letter alphabets;
  Let $\alpha(a) = ad$, $\alpha(b) = ac$, $\alpha(c) = cb$ and $\alpha(d) = ca$, while $\delta(a) = fgifh$, $\delta(b) = fghij$, $\delta(c) = jigjh$ and $\delta(d) = jihgf$. Take $\delta(\alpha^\omega(a))$
- avoid all overlaps, over five letter alphabets;
  (an overlap is of the form $axbyx$ with $axb \uparrow byc$)
  For $\beta(a) = defghijk$, $\beta(b) = deghfkij$ and $\beta(c) = dehfgjki$, take $\beta(\sigma)$
- avoid all cubes, over four letter alphabets;
  For $\gamma(a) = abcd$ and $\gamma(b) = badc$, take $\alpha(\tau)$

[Blanchet-Sadri, M., Scott, TCS 09] [Manea, M., TCS 07]

Algorithms

Let w be a word of length n, and k, p and d some positive integers.

One can identify as fast as $\mathcal{O}(\frac{n^2}{k})$ if the word w is k-free.

One can identify as fast as $\mathcal{O}(n \log(n))$ all periods of the word w.

[Manea, M., TCS 07] [Manea, M., Tiseanu, MFCS 2011]

Periodicity

One can identify as fast as $\mathcal{O}(n)$, if inserting holes into the word w, such that between each two holes there are at least $d-1$ letters of the alphabet, gives us a p-periodic partial word.

One can preprocess a word as fast as $\mathcal{O}(n \log(n))$ time, such that it can answer in constant time queries of the form:

- "Is the factor $w[i..j]$ p-periodic?"
- "Which is the minimum period of the factor $w[i..j]$?"

Furthermore, we continue answering the above queries in constant time after an $\mathcal{O}(n)$ update taking place during an online extension of our word.

[Manea, M., Tiseanu, MFCS 2011]

Avoidable patterns

A word w avoids a pattern p if for no non-decreasing function f from the alphabet of p to the alphabet of w, we have $f(p)$ as a factor of w.

The avoidable binary patterns are fully characterized. When you consider non-trivial avoidability the results are the same as in the full word case

For ternary alphabets only 4 patterns are missing (*ABACBC*, *AABCCAB*, *AABACABBA*, and *ABBACAABA*).

[Blanchet-Sadri, M., Simmons, Weissenstein, Acta Informatica 11]
[Blanchet-Sadri, Black, Zemke, LATA 11]

Large squares

All infinite binary partial words whose factors are compatible with no more than three distinct full squares have at most two holes. There exists a binary word with infinitely many holes whose factors are compatible only with the full squares aa, bb, $abab$ and $bbbb$.

All infinite cube-free binary partial words whose factors are compatible with no more than ten distinct full squares have at most two holes. There exists a cube-free binary word with infinitely many holes whose factors are compatible only with the eleven full squares a^2, b^2, $(ab)^2$, $(ba)^2$, $(aab)^2$, $(aba)^2$, $(abb)^2$, $(baa)^2$, $(bab)^2$, $(bba)^2$, and $(abba)^2$.

[Blanchet-Sadri, Choi, M., TCS 11]

Repetition threshold

Which is the repetition threshold for a n-letter alphabet?

RT(n)=inf$\{k \mid$ an infinite word over an n-letter alphabet avoids k-powers.$\}$
Dejean conjectured that for positive integer $n > 1$,

$$RT(n) = \begin{cases} 7/4 & n = 3 \\ 7/5 & n = 4 \\ n/(n-1) & n \notin \{3, 4\} \end{cases}$$

The threshold for a two letter alphabet is $\frac{5}{2}$, while for alphabets larger than two, the threshold is 2.

[Halava, Harju, Kärki, Séébold, TCS 09]

Counting Squares

Which is the maximum number of distinct squares a word of length n can have as factors?

It is proved that the number is less than $2n - \Theta(\log n)$, but conjectured that is n (recently it is conjectured for binary words that this is $\frac{2k-1}{2k+2}$, where k is the number of occurrences of the second letter)

The maximum number of distinct squares compatible with the factors of a partial word with one hole is less than $\frac{7n}{2}$. If the alphabet is binary then the bound is $3n$.

[Blanchet-Sadri, M., Scott, Acta Cybern. 09] [Blanchet-Sadri, M., ITA 09] [Halava, Harju, Kärki, ITA 10]

Hole sparsity

How close can we have holes within a word that still avoids repetitions?

If we disconsider the case when one instance of α could be a hole. The avoidability of unary patterns.

k	α	α^2	α^3	α^4	α^5	$\cdots$
1	∞	∞	∞	∞	∞	$\cdots$
2	∞	∞	3	2	2	$\cdots$
3	∞	7	2	2	2	$\cdots$
4	$\cdots$	4	2	2	2	$\cdots$
5	$\cdots$	4	2	2	2	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

k	α	α^2	α^3	α^4	α^5	α^6	α^7	$\cdots$
1	∞	∞	∞	∞	∞	∞	∞	$\cdots$
2	∞	∞	7	3	3	2	2	$\cdots$

Abelian repetitions

Two words are abelian equivalent if one is a permutation of the other.

intestines

Over a 5-letter alphabet there exist infinite words that contain no factors of length greater than 2, which are compatible with some abelian square.

Over a 4-letter alphabet there exist infinite words that contain no factors which are compatible with some abelian cube. Over a 3-letter alphabet

k-abelian complexity

Two words are k-abelian equivalent if the multisets given by all factors of length at most k of each, are equal.

$$aba \not\equiv_2 aab$$

$$baabba \equiv_2 babbaa$$

Using a morphism designed to get results regarding the hole sparsity of cubes in partial words, the next result was obtained:

Cubes are 4-abelian avoidable over a binary alphabet.

[M., Saarela, Journees Montoises 12]

Avoidability index

Cassaigne conjectured in 1994 that any pattern with m distinct variables of length at least $3(2^{m-1})$ is avoidable over a binary alphabet, and any pattern with m distinct variables of length at least 2^m is avoidable over a ternary alphabet.

[Blanchet-Sadri, Woodhouse, DLT 13] [Ochem, Pinlou, CoRR 13]

Mulțumesc

Bibliography I

- [Levitin 07] Daniel Levitin: This is Your Brain on Music: The Science of a Human Obsession. Plume (2007)
- [Manea, M., TCS 07] Florin Manea, Robert Mercaș: Freeness for partial words. Theoret. Comput. Sci. **389** 1-2 (2007) 265–277
- [Blanchet-Sadri, M., Scott, TCS 09] Francine Blanchet-Sadri, Robert Mercaș, Geoffrey Scott: A generalization of Thue freeness for partial words Theoret. Comput. Sci. **410** 8-10 (2009) 793–800
- [Blanchet-Sadri, M., Simmons, Weissenstein, Acta Informatica 11] Francine Blanchet-Sadri, Robert Mercaș, Sean Simmons, Eric Weissenstein: Avoidable binary patterns in partial words. Acta Informatica **48** 1 (2011) 25–41
- [Blanchet-Sadri, Black, Zemke, LATA 11] Francine Blanchet-Sadri, Kevin Black, Andrew Zemke: Unary Pattern Avoidance in Partial Words Dense with Holes. LATA (2011) 155-166

Bibliography II

- [Blanchet-Sadri, Choi, M., TCS 11] Francine Blanchet-Sadri, Ilkyoo Choi, Robert Mercas: Avoiding large squares in partial words. Theor. Comput. Sci. **412** 29 (2011) 3752–3758
- [Halava, Harju, Kärki, Séébold, TCS 09] Vesa Halava, Tero Harju, Tomi Kärki, Patrice Séébold Overlap-freeness in infinite partial words. Theor. Comput. Sci. **410** 8-10 (2009) 943–948
- [Blanchet-Sadri, M., Scott, Acta Cybern. 09] Francine Blanchet-Sadri, Robert Mercas, Geoffrey Scott: Counting Distinct Squares in Partial Words. Acta Cybernetica **19** 2 (2009) 465–477
- [Blanchet-Sadri, M., ITA 09] Francine Blanchet-Sadri, Robert Mercas: A note on the number of squares in a partial word with one hole. ITA **43** 4 (2009) 767–774
- [Halava, Harju, Kärki ITA 10] Vesa Halava, Tero Harju, Tomi Kärki: On the number of squares in partial words. ITA **44** 1 (2010) 125–138

Bibliography III


[Blanchet-Sadri, Lohr, Scott, IWOCA 12] Francine Blanchet-Sadri, Andrew Lohr, Shane Scott: Computing the Partial Word Avoidability Indices of Ternary Patterns. IWOCA (2012) 206–218


[M., Saarela, Journees Montoises 12] Robert Mercas, Aleksi Saarela: 5-Abelian Cubes Are Avoidable on Binary Alphabet. Journees Montoises (2012)


[Blanchet-Sadri, Woodhouse, DLT 13] Francine Blanchet-Sadri, Brent Woodhouse: Strict Bounds for Pattern Avoidance. Developments in Language Theory (2013) 106–117


[Manea, M., Tiseanu, MFCS 2011] Florin Manea, Robert Mercas, Catalin Tiseanu: Periodicity Algorithms for Partial Words. MFCS (2011) 472–484


[Ochem, Pinlou, CoRR 13] Pascal Ochem, Alexandre Pinlou: Application of entropy compression in pattern avoidance. CoRR abs/1301.1873 (2013)